Michał Bernardelli

Szkoła Główna Handlowa w Warszawie
Kolegium Analiz Ekonomicznych

michal.bernardelli@sgh.waw.pl

Modelowanie danych panelowych z wykorzystaniem obliczeń rozproszonych w Apache Spark

Apache Spark to nowy, bo powstały w roku 2009, otwarty framework do rozpraszania obliczeń na klastrach komputerowych. Stale rozwijany i udoskonalany, w najnowszej wersji 1.6.0 oferuje możliwości efektywnej analizy wielkich wolumenów danych, dzięki stosunkowo rozbudowanej bibliotece MLlib (Machine Learning Library). W skład biblioteki wchodzą przede wszystkim iteracyjne algorytmy uczenia maszynowego, służące przede wszystkim zagadnieniom klasyfikacji oraz regresji. Mimo dostępnych implementacji różnych rodzajów regresji, np. liniowej, logistycznej, z uwzględnieniem regularyzacji, nie można tych metod zastosować bezpośrednio do analizy danych panelowych.

Celem artykułu jest przedstawienie sposobu wyznaczania parametrów modeli opartych na danych panelowych z wykorzystaniem tej potężnej biblioteki. Dzięki wprowadzeniu dodatkowych zmiennych, zastosowaniu twierdzenia Frischa-Waugha oraz modelu programistycznego Map-Reduce, stworzona została efektywna obliczeniowo metoda wyznaczania estymatora LSDV (least squares dummy variables) w Apache Spark. Efektywność obliczeń została wykazana poprzez przeprowadzenie symulacji komputerowych.